CONSISTENCY OF BAYESIAN INFERENCE OF RESOLVED PHYLOGENETIC TREES

MIKE STEEL

Abstract. Bayesian methods are now one of the main tools for inferring phylogenetic trees from aligned sequence data. A recent paper [3] has suggested that Bayesian tree inference might be statistically inconsistent, at least for certain combinations of branch lengths (in part of the 'Felsenstein Zone'), on the basis of their simulation results. We mathematically demonstrate that Bayesian methods for inferring resolved trees are statistically consistent over the entire parameter range of branch lengths, under general conditions on priors, at least for simple models where the usual 'identifiability' conditions hold.

1. Introduction

Bayesian Inference (BI) has become a mainstream approach for inferring phylogenetic tree topology from aligned DNA sequence data [4]. The approach has a number of desirable features, however some potential problems with the convergence of Bayesian methods (MCMC) have been highlighted [5]. The paper [3] investigated the performance of Bayesian inference (BI) for inferring phylogenetic tree topology from aligned DNA sequence data under various models. Among the interesting findings in their simulation study was that Bayesian methods applied to unresolved 4–leaf trees (with a zero-length interior edge) with certain combinations of long/short pendant branches tended to show increasing bias towards one of three particular resolved trees as the sequence length increased, in contrast to maximum likelihood which favored each of the three resolutions equally (*c.f.* [8]).

The authors suggested the possibility that for data generated by a resolved tree, with a certain combination of short and long edges, BI might even be statistically inconsistent (i.e. the tree with the highest posterior probability for the data being different to the tree that generated the data, with a probability that does not tend to zero as the sequence length grows) even for models for which maximum likelihood is known to be statistically consistent [1]. Noting that "our finding provide strong, albeit circumstantial evidence that BI is statistically inconsistent," the authors of [3] were careful to point out that a formal mathematical analysis of this possibility was beyond the scope of their paper.

In this short paper we provide a self-contained formal proof that BI is statistically consistent for inferring the topology of resolved phylogenetic trees over the entire range of branch lengths, and under essentially the same identifiability conditions that establish the consistency of maximum likelihood estimation of tree topology. We do this by establishing a more general result that includes the phylogenetic setting as a particular case.

1.1. General result.

Consider the general problem of identifying a discrete parameter lying in an arbitrary finite set A from a sequence of i.i.d. observations that take values in an arbitrary finite set U. Suppose further that the probability distribution on U is determined not just by the discrete parameter $a \in A$, but by some additional ('nuisance') parameters. In this paper we will assume in that these additional parameters are continuous, and if we denote the parameter

I thank Joe Thornton for clarifying some comments in [3].

space associated with each discrete parameter $a \in A$ by $\Theta(a)$ that this is an open subset of some Euclidean space.

In the phylogenetic setting, A is the set of fully resolved (binary) phylogenetic tree topologies on a given leaf set, U is the set of possible site patterns, and the parameter set $\Theta(a)$ specifies for the tree topology a all its possible branch lengths, each of which lies in the range $(0, \infty)$. Thus $\Theta(a) = (0, \infty)^{2n-3}$ where n is the number of leaves of tree a.

Let $p_{(a,\theta)}$ denote the probability distribution on some finite set U determined by the discrete-continuous parameter pair (a, θ). Suppose we have a discrete (prior) probability distribution π on A, and, for each $a \in A$, a continuous (prior) probability distribution on $\Theta(a)$ with probability density function $f_a(\theta)$. We will suppose that the following conditions hold for all $a \in A$:

(C1) $\pi(a) > 0$;
(C2) The density $f_a(\theta)$ is a continuous, bounded and nonzero on $\Theta(a)$;
(C3) The function $\theta \mapsto p_{(a,\theta)}$ is continuous and nonzero on $\Theta(a)$;
(C4) For all $\theta \in \Theta(a)$, and all $b \neq a$, we have: $\inf_{\theta' \in \Theta(b)} d(p_{(a,\theta)}, p_{(b,\theta')}) > 0$,

where in (C4) and henceforth, d denotes the L_1 metric – that is, for any two probability distributions p, q on U:

$$d(p, q) := \sum_{u \in U} |p(u) - q(u)|.$$

In the phylogenetic setting, if π is any of the usual non-zero priors on binary phylogenetic trees (e.g. the uniform (PDA) distribution, or the Yule distribution), then condition (C1) is satisfied. If we take the usual exponential prior on branch lengths then condition (C2) is satisfied. For all Markov processes on trees condition (C3) holds (the nonzero condition holds since in any tree with positive length pendant edges all site patterns have a strictly positive probability). Finally, for all models for which identifiability holds (eg. the GTR model or any submodel down to the most restrictive the Jukes-Cantor 1969 model) condition (C4) holds (see e.g. [7]; a specific lower bound on d for the 2-state symmetric model is provided via Lemma 7.3 of [6]).

Now, suppose we are given a sequence $\mathbf{u} = (u_1, \ldots, u_k) \in U^k$ generated i.i.d. by some unknown pair (a, θ) and we wish to identify the discrete parameter (a) from $\mathbf{u}$ given prior densities on A and the continuous parameters. The *Maximum a-Posteriori* (MAP) estimator selects the element $b \in A$ that maximizes the posterior probability of b given $\mathbf{u}$ – that is, it maximizes $\pi(b)\mathbb{E}_{\theta'}[\mathbb{P}(\mathbf{u}|b, \theta')]$, where:

$$\mathbb{P}(\mathbf{u}|b, \theta) = \prod_{j=1}^{k} p_{(b,\theta')}(u_i), \tag{1}$$

is the probability of generating the sequence of i.i.d. observations $(u_1, \ldots, u_k)$ from underlying parameters (b, θ'), and where $\mathbb{E}_{\theta'}$ refers to expectation with respect to the prior probability distribution on $\Theta(b)$.

Let $P(a, \theta, k)$ denote the probability that, for a sequence $u_1, \ldots, u_k$ generated i.i.d. by (a, θ), the MAP estimator correctly selects a. The following theorem establishes a sufficient condition for the statistical consistency of the MAP estimator in this context.

Theorem 1. *Provided conditions (C1)–(C4) hold for all $a \in A$, then $\lim_{k\to\infty} P(a, \theta, k) = 1$ for all $a \in A$, and $\theta \in \Theta(a)$.*

2. Proof of Theorem 1

We first present a lemma that is helpful in the proof of Theorem 1.

Lemma 2. *For any $\epsilon_1, \epsilon_2, \epsilon_3 > 0$ there exists $\delta > 0$ and $\epsilon > 0$ so that the following holds: For any finite set U, and probability distributions p, q, r, s on U that satisfy the three conditions:*

(i) $d(p, q) \geq \epsilon_1$;
(ii) $p(u) \geq \epsilon_2$, $q(u) \geq \epsilon_3$ *for all* $u \in U$;
(iii) $d(p, r) < \delta$ *and* $d(p, s) < \delta$;

we have:

$$\sum_{u \in U} r(u) \log\left(\frac{s(u)}{q(u)}\right) \geq \epsilon. \tag{2}$$

Proof. We will show that any $\delta > 0$ and $\epsilon > 0$ for which $f(\epsilon, \delta) \geq \epsilon > 0$ suffices for (2), where:

$$f(\epsilon, \delta) := \frac{1}{2}(\epsilon_1 - \delta)^2 - 2\delta \cdot (|\log(\epsilon_2 - \delta)| + |\log(\epsilon_3)|).$$

For each $u \in U$, let $\Delta_u := r(u) - s(u)$. Then:

$$\sum_{u \in U} r(u) \log\left(\frac{s(u)}{q(u)}\right) = \sum_{u \in U} s(u) \log\left(\frac{s(u)}{q(u)}\right) + \sum_{u \in U} \Delta_u \log\left(\frac{s(u)}{q(u)}\right). \tag{3}$$

Now, the first term on the right hand-side of (3) is simply the Kullback-Leibler separation of s and q and, by Pinsker's Inequality [2], this is bounded below by $\frac{1}{2}d(s, q)^2$. Moreover, by the triangle inequality, $d(s, q) \geq (p, q) - d(s, p) \geq \epsilon_1 - \delta$ (by conditions (i) and (iii)), and so the first term on the right of (3) is bounded below by $\frac{1}{2}(\epsilon_1 - \delta)^2$.

Concerning the second term on the right of (3), its absolute value is bounded above by:

$$\sum_{u \in U} |\Delta_u| \max_{u \in U} |\log\left(\frac{s(u)}{q(u)}\right)| = d(r, s) \cdot \max_{u \in U} |\log\left(\frac{s(u)}{q(u)}\right)|.$$

Again invoking the triangle inequality, $d(r, s) \leq d(r, p) + d(p, s) \leq 2\delta$ (by condition (iii)); moreover, since $s(u) \geq p(u) - \delta \geq \epsilon_2 - \delta$, and $q(u) \geq \epsilon_3$ (both by condition (iii)):

$$\max_{u \in U} |\log\left(\frac{s(u)}{q(u)}\right)| \leq \max_{u \in U} |\log(s(u))| + \max_{u \in U} |\log(q(u))| \leq |\log(\epsilon_2 - \delta)| + |\log(\epsilon_3)|.$$

Applying this (and the earlier) bound to (3) gives: $\sum_{u \in U} r(u) \log\left(\frac{s(u)}{q(u)}\right) \geq f(\epsilon, \delta)$, as required. □

Turning to the proof of Theorem 1, we suppose throughout that the sequence $\mathbf{u} = u_1, \ldots, u_k$ is generated i.i.d. by (a, θ_0) where θ_0 is any particular element of $\Theta(a)$. Then the MAP estimator will correctly select a from $\mathbf{u}$ if and only if the *Bayes Factor* defined by:

$$BF_{a/b} = \frac{\pi(a)\mathbb{E}_\theta[\mathbb{P}(\mathbf{u}|a, \theta)]}{\pi(b)\mathbb{E}_{\theta'}[\mathbb{P}(\mathbf{u}|b, \theta')]}$$

is strictly greater than 1 for all $b \neq a$. By the Bonferroni inequality, it suffices to show that for each $b \neq a$ the probability that $\mathbf{u}$ is such that $BF_{a/b} > 1$ tends to 1 as k grows. To achieve this we first observe that $BF_{a/b} = \frac{\pi(a)}{\pi(b)} \cdot R_{a/b}$ where:

$$R_{a/b} := \frac{\mathbb{E}_\theta[\mathbb{P}(\mathbf{u}|a, \theta)]}{\mathbb{E}_{\theta'}[\mathbb{P}(\mathbf{u}|b, \theta')]}, \tag{4}$$

and where $\frac{\pi(a)}{\pi(b)}$, is a strictly positive by (C1). Thus it suffices to show that, for each $b \neq a$ and for any finite constant M, the inequality $R_{a/b} > M$ holds with a probability that tends to 1 as

k tends to infinity. We will establish this inequality by providing an explicit lower bound to the numerator of $R_{a/b}$ and an explicit upper bound to the denominator of $R_{a/b}$, and showing that, with probability tending to 1 as k grows, their ratio exceeds M.

Before describing the lower bound, observe that we can re-write Eqn. (1) as follows:

$$\mathbb{P}(\mathbf{u}|b,\theta) = \prod_{u\in U} p_{(b,\theta)}(u)^{n_u} = \prod_{u\in U} p_{(b,\theta)}(u)^{r(u)k}, \tag{5}$$

where, for each $u \in U$, $n_u := |\{j : u_j = u\}|$, and $r(u) := \frac{1}{k}n_u$. Note that $r = (r(u) : u \in U)$ is a (empirical) probability distribution on U (i.e. its entries are nonzero and sum to 1). In the phylogenetic setting r describes the frequency of site patterns in the data.

For the lower bound on the numerator of $R_{a/b}$, consider the subset N_τ of $\Theta(a)$ consisting of a closed ball centered on θ_0 and of radius $\tau > 0$. Note that we can always select a sufficiently small value of $\tau > 0$ for which $N_\tau \subset \Theta(a)$ by the assumption that $\Theta(a)$ is an open subset of some Euclidean space. Letting $\mu(N_\tau) = \int_{N_\tau} f_a(\theta)d\theta > 0$ we have:

$$\mathbb{E}_\theta[\mathbb{P}(\mathbf{u}|a,\theta)] = \int_{\Theta(a)} \mathbb{P}(\mathbf{u}|a,\theta)f_a(\theta)d\theta \geq \int_{N_\tau} \mathbb{P}(\mathbf{u}|a,\theta)f_a(\theta)d\theta \geq \mu(N_\tau)\cdot \inf_{\theta\in N_\tau}\{\mathbb{P}(\mathbf{u}|a,\theta)\}. \tag{6}$$

Now, let s be the probability distribution of the form $p_{(a,\theta)}$ that minimizes $\mathbb{P}(\mathbf{u}|a,\theta)$ when θ is restricted to N_τ; such a distribution s exists from the compactness of N_τ and the continuity condition of (C3). Then, from (5) we have: $\inf_{\theta\in N_\tau}\{\mathbb{P}(\mathbf{u}|a,\theta)\} = \prod_{u\in U} s(u)^{r(u)k}$. Applying this to (6) gives:

$$\mathbb{E}_\theta[\mathbb{P}(\mathbf{u}|a,\theta)] \geq \mu(N_\tau)\cdot \prod_{u\in U} s(u)^{r(u)k}. \tag{7}$$

Regarding the upper bound on the denominator of $R_{a/b}$, we have:

$$\mathbb{E}_{\theta'}[\mathbb{P}(\mathbf{u}|b,\theta')] \leq \sup_{\theta'\in\Theta(b)}\{\mathbb{P}(\mathbf{u}|b,\theta')\}. \tag{8}$$

Given $\mathbf{u}$, let θ_i be a sequence of elements of $\Theta(b)$ for which $\lim_{i\to\infty}\mathbb{P}(\mathbf{u}|b,\theta_i) = \sup_{\theta'\in\Theta(b)}\{\mathbb{P}(\mathbf{u}|b,\theta')\}$. Then $p_{(b,\theta_i)}$, $i \geq 1$, is a sequence in a bounded subset of Euclidean space (the probability simplex) and so, by the Bolzano–Weierstrass theorem, it has a convergent subsequence, with limit q (a probability distribution on U).

From Eqns. (4), (7) and (8) we have:

$$R_{a/b} \geq \frac{\mu(N)\cdot\prod_{u\in U} s(u)^{r(u)k}}{\prod_{u\in U} q(u)^{r(u)k}}$$

and so:

$$\log(R_{a/b}) \geq \log(\mu(N_\tau)) + k\sum_{u\in U} r(u)\log\left(\frac{s(u)}{q(u)}\right). \tag{9}$$

In particular, for any positive constant $M > 1$, we have $R_{a/b} > M$ as $k \to \infty$ provided that, for some $\epsilon > 0$ that does not depend on k or τ:

$$\sum_{u\in U} r(u)\log\left(\frac{s(u)}{q(u)}\right) \geq \epsilon > 0. \tag{10}$$

Definitions: For $\rho > 0$ let E_ρ be the event that $\mathbf{u}$ is such that $d(p,r) < \rho$, where r is the empirical distribution determined by $\mathbf{u}$ and defined after Eqn. (5). Let $p := p_{(a,\theta_0)}$, and let $\epsilon_2 := \min\{p(u) : u \in U\} > 0$ (by (C3)).

Claim: For $\rho = \frac{1}{2}\epsilon_2$, if $\mathbf{u}$ satisfies E_ρ then $q(u) \geq \epsilon_3$ for some $\epsilon_3 > 0$ determined by ϵ_2.

Proof of Claim. By condition (C3) (applied to b) if we select any $\theta_1 \in \Theta(b)$ then for some $\nu > 0$ we have $p_{(b,\theta_1)}(u) \geq \nu$ for all $u \in U$, and so:

$$\mathbb{P}(\mathbf{u}|b, \theta_1) \geq \nu^k. \tag{11}$$

Now, if $q(u) < \epsilon_3$ for at least one $u \in U$, then $\sup_{\theta' \in \Theta(b)}\{\mathbb{P}(\mathbf{u}|b, \theta')\} \leq q(u)^{n_u} < \epsilon_3^{n_u}$. Moreover, if $\mathbf{u}$ satisfies E_ρ, then $n_u \geq k(p(u) - \rho)$ and, since $p(u) \geq \epsilon$ and $\rho = \epsilon/2$ we have:

$$\sup_{\theta' \in \Theta(b)} \{\mathbb{P}(\mathbf{u}|b, \theta')\} < \epsilon_3^{k(p(u)-\rho)} \leq \epsilon_3^{k\epsilon_2/2}. \tag{12}$$

However if ϵ_3 is small enough that: $\epsilon_3^{\epsilon_2/2} < \nu$ then (11) and (12) imply $\mathbb{P}(\mathbf{u}|b, \theta_1)$ is strictly greater than $\sup_{\theta' \in \Theta(b)}\{\mathbb{P}(\mathbf{u}|b, \theta')\}$, a contradiction. Thus for such a small value of ϵ_3 we have $q(u) \geq \epsilon_3$ for all $u \in U$, which establishes the Claim.

Returning to the proof of Theorem 1, we are now in a position to apply Lemma 2. For a fixed $b \neq a$, let $\epsilon_1 = \inf_{\theta' \in \Theta(b)}\{d(p, p_{(b,\theta')})\}$. By (C4), $\epsilon_1 > 0$, and since q is either of the form $p_{(b,\theta')}$, or is the limit of such distributions as $\theta' \in \Theta(b)$ varies, we have:

$$d(p, q) \geq \epsilon_1 > 0.$$

Recall that $\epsilon_2 = \min\{p(u) : u \in U\} > 0$ by (C2). Set $\rho = \frac{1}{2}\epsilon_2$, so that if $\mathbf{u}$ satisfies E_ρ then (by the Claim), $q(u) \geq \epsilon_3 > 0$. Applying Lemma 2 there exists $\epsilon > 0$ and $\delta > 0$ (dependent just on $\epsilon_1, \epsilon_2, \epsilon_3$) so that, provided $\mathbf{u}$ satisfies the event $E_{\epsilon_2/2}$ (in order for condition (iii) to hold for q), and provided $d(p, r), d(p, s) \leq \delta$, then Inequality (10) holds. Now, the event that $d(p, r) \leq \delta$ is precisely the requirement $\mathbf{u}$ satisfies event E_δ, and for any $\delta > 0$ this event has probability converging to 1 as k grows, by the law of large numbers (the same comment applies for the requirement that $\mathbf{u}$ satisfies $E_{\epsilon_2/2}$). Moreover, by the continuity condition in (C3), we can select $\tau > 0$ sufficiently small so that $d(p, s) < \delta$. Although a small τ inflates the magnitude of the negative first term $(\log(N_\tau))$ on the right of (9), the second term is positive (by (10)) and grows linearly with k. Thus, for any finite value M, $R_{a/b} > M$ with probability converging to 1 as $k \to \infty$. By the comments following (4), this completes the proof. □.

Remark In the proof, notice that only the probability distribution $p = p_{(a,\theta_0)}$ is fixed, the other three distributions r, s, q depend on the data $\mathbf{u}$ that is generated by p. However, Lemma 2 quantifies over all choices of r, s, q once the ϵ_i values have been specified, and these, in turn, depend ultimately just on p.

References

[1] Chang, J.T. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math. Biosci. 137, 5173.

[2] Cover, T.M., Thomas, J.A. 1991. Elements of Information Theory. John Wiley & Sons, Inc., New York.

[3] Kolackzkowski, B., Thornton, J. W. 2009. Long-branch attraction bias and inconsistency in Bayesian phylogenetics. PLos One, 4 (12) e7891.

[4] Lemey, P., Salemi, M., Vandamme, A.-M. 2009. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing (2nd Ed.). Cambridge University Press.

[5] Mossel, E., Vigoda, E. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science 309 2207–2209.

[6] Steel, M., L. Székely. 2007. Teasing apart two trees. Combinatorics, Probability and Computing 16, 903-922.

[7] Steel, M., L. Székely. 2009. Inverting random functions (III): Discrete MLE revisited. Annals of Combinatorics 13: 373390.

[8] Susko, E. 2008. On the distributions of bootstrap support and posterior distributions for a star tree. Systematic Biology. 57: 602 612.

Mike Steel: Allan Wilson Centre for Molecular Ecology and Evolution, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

E-mail address: m.steel@math.canterbury.ac.nz